VALIDATION OF QUALITY RATING AND IMPROVEMENT SYSTEMS FOR EARLY CARE AND EDUCATION AND SCHOOL-AGE CARE

Research-to-Policy, Research-to-Practice Brief OPRE2012-29
April 2012

OPRE
Office of Planning, Research and Evaluation
Administration for Children and Families

DISCLAIMER:

The views expressed in this publication do not necessarily represent the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families or the U.S. Department of Health and Human Services.

ACKNOWLEDGMENTS

The authors would like to thank Ivelisse Martinez-Beck and Naomi Goldstein at the Office of Planning, Research and Evaluation, Kathryn Tout at Child Trends, and Laura Hamilton at RAND for their guidance and feedback on this paper.

Validation of Quality Rating and Improvement Systems for Early Care and Education and School-age Care

Research-to-Policy, Research-to-Practice Brief OPRE2012-29

April 2012

Submitted to:
Ivelisse Martinez-Beck, PhD., Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Submitted by:
Gail L. Zellman, RAND Corporation
Richard Fiene, Pennsylvania State University

Contract Number: GS10F0030R
Project Director: Kathryn Tout
Child Trends
4301 Connecticut Ave NW
Washington DC, 20008

Suggested Citation:
Zellman, G. L. & Fiene, R. (2012). *Validation of Quality Rating and Improvement Systems for Early Care and Education and School-Age Care,* Research-to-Policy, Research-to-Practice Brief OPRE 2012-29. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This Brief was developed by members of the Quality Initiatives Research and Evaluation Consortium (INQUIRE) which is designed to facilitate the identification of issues and the development and exchange of information and resources related to research and evaluation of quality rating and improvement systems (QRIS) and other quality initiatives. INQUIRE is funded by the Office of Planning, Research and Evaluation through the Child Care and Early Education Policy and Research Analysis and Technical Expertise contract with Child Trends.

Validation of Quality Rating and Improvement Systems for Early Care and Education and School-age Care

Quality Rating and Improvement Systems (QRIS) for early care and education and school age care programs are designed to collect information about quality and to use that information to produce program-level ratings, which are the foundation of a QRIS. The ratings are intended to make program quality transparent for parents and other stakeholders and to encourage the selection of higher-quality programs. The ratings also provide benchmarks that can support efforts to help programs improve their quality. ***Validation* of a QRIS is a multi-step process that assesses the degree to which design decisions about *program quality standards and measurement strategies* are resulting in accurate and meaningful ratings.** Validation of a QRIS provides designers, administrators and stakeholders with crucial data about how well the architecture of the system is functioning. A carefully designed plan for ongoing validation creates a climate that supports continuous quality improvement at both the program and system level.

To date, QRIS validation efforts have been limited. One reason may be that validation is a complex endeavor that involves a range of activities. In addition, there has been little guidance available that clarifies the purpose of QRIS validation or identifies the activities that comprise validation. At the same time, there is growing pressure to validate these systems as stakeholders seek evidence that QRIS are functioning as intended. The federal government has elevated QRIS validation by including it as a central component of the 2011 Race to the Top Early Learning Challenge and requiring state applicants to develop QRIS validation plans as part of their submissions.

The purpose of this Brief is to help QRIS stakeholders better understand validation and to outline a set of complementary validation activities. The Brief defines validation, describes different types of validation studies, and provides guidance on developing a validation plan, including tools to determine the appropriate scope and timing of validation activities. It also lists references and resources for those who wish to learn more. This Brief is aimed at readers in positions to authorize, finance, design, and refine QRISs and other quality improvement efforts, including state child care administrators, early education policy and program specialists, legislators, and other potential funders.

QRIS Validation and Its Role in Continuous System Improvement

***Validation* is a multi-step process that assesses the degree to which design decisions about QRIS *program quality standards and measurement strategies* are resulting in accurate and meaningful program ratings.**[1]

Validation is particularly important for QRISs because these systems at their core rely on ratings of program quality. They are built on the assumption that the quality of early childhood and school-age programs can be reliably measured and that differences in quality across these programs can be identified through the use of a set of quality indicators. Validity data can support conclusions about whether such quality indicators measure quality well and whether the strategies used to combine measures and develop ratings are working as intended (Cizek, 2007).[2] Valid ratings are critical to QRISs because parents and other stakeholders use these ratings to select the highest-quality care that they can afford. The overall quality rating also carries increasingly high stakes for programs. Indeed, the theory underlying QRISs intentionally creates those stakes to motivate both provider and parent behaviors in support of increased quality (e.g., Zellman et al., 2008; Zellman et al., 2011). In addition to attracting more children, programs that score well may receive higher subsidies for subsidy-eligible children, and may qualify for grants, incentives, and tax credits.

Why QRIS validation is important. A QRIS is a primary strategy states employ to improve early childhood education and school-age care (ECE-SAC) program quality. Because ratings are a central element of a QRIS, it is important to collect data to establish that these ratings are accurate and meaningful indicators of quality. Validation studies can lend credibility to a QRIS, identify needed changes, and support continuous improvement of a QRIS.

Validity is not determined by a single study; instead, validation should be viewed as a continuous process with multiple goals: refining the ratings, improving system functioning, and increasing the credibility and value of rating outcomes and of the QRIS system as a whole. A carefully designed validation plan will promote the accumulation of evidence over time that will provide a sound theoretical and empirical basis for the QRIS (AERA, APA, & NCME, 1999; Kane, 2001). Ongoing validation activities that are carried out in tandem with QRIS monitoring activities (that aim to examine ongoing implementation of the QRIS) and evaluation activities (that examine the outcomes of QRIS) can help a QRIS improve its measures and effectiveness throughout its development and implementation (see Lugo-Gil et al., 2011 and Zellman et al., 2011 for guidance on developing a comprehensive QRIS evaluation).

[1] The definition of validation has changed over time. Rather than identifying separate types of validity (construct, predictive, face, concurrent and content), the current notion is that construct validity includes all evidence for validity, including content and criterion evidence, reliability, and the wide range of methods associated with theory testing (Messick, 1975, 1980; Tenopyr, 1977; Guion, 1977; Embretson, 1983; Anastasi, 1986). As a consequence, we do not differentiate types of validity in this brief.

[2] Reliability represents the ability of a measure to assess its target behaviors or characteristics consistently. In the case of QRISs, reliability refers to the extent to which independent raters produce similar ratings on individual QRIS elements and on the summary rating (inter-rater reliability) as well as the degree to which raters are consistent over time in their ratings (intra-rater reliability). Such consistency is a prerequisite for validity of any measure.

QRIS validation activities may produce three important benefits. First, validation evidence can promote increased support for the system among parents, ECE-SAC providers and other key stakeholders. Ratings that match the experiences of parents and providers can build trust in the ratings and increase the overall credibility of the system. Second, a system that is measuring quality accurately is better able to target limited quality improvement supports to those programs and program elements most in need of improvement. Third, validation evidence can be used to improve the efficiency of the rating process. If a QRIS is expending resources to measure a component of quality that is not making a unique contribution to a summary quality rating or that is not measuring quality accurately, it can be removed or revised. For example, measures that vary little if at all across providers whose quality varies substantially in other ways make little or no contribution to quality ratings. Measures of family engagement that include parent ratings are particularly prone to this problem, as parents who have chosen to use and continue to rely on a given provider are highly likely to see the care as good and to rate it according to their views (Zellman and Perlman, 2006; McGrath, 2007; Keyes, 2002; Kontos et al., 1987; Shimoni, 1992). If all or almost all programs receive high ratings on the family engagement measure, then that component of the rating may not be working to distinguish between lower-quality and higher-quality programs. It may be considered important to collect measures of family engagement to ensure that providers continue to focus on it. But knowing that a given measure is not contributing to an overall program quality rating may motivate program developers to consider another way to measure the concept, which might both increase the value of the measure and reduce measurement costs. Indeed, understanding the relationships among rating elements through validation studies can save substantial time and effort.

Despite the importance of validation activities to strengthen QRIS, support for these activities may be impeded by limited resources and concern about the value of validation activities. In states with more mature QRISs, there may be reluctance among stakeholders to assess an established system. In newer systems, policymakers may question the need for validation given the arguments recently offered in support of establishing the system. Validation plans can address each of these concerns by providing evidence to help the system run more efficiently and to establish a climate of continuous improvement. A validation plan will clarify that the system is open to change, intent on improvement, and dedicated to increasing the odds of reaching its goals.

Designing and Implementing Validation Efforts

A comprehensive validation plan includes multiple studies that rely on different sources of information and ask different but related questions. These can be understood and organized around four complementary and interrelated approaches to validation. In this section we provide details of the four approaches. Summaries of these details are provided in two tables. Table 1 presents an overview of the four approaches including the purpose of each approach, the activities that might be undertaken, the questions that are asked and the limitations of each approach. Table 2 presents the data needed, data sources, and analysis methods for selected studies within each approach.[3]

[3] The four basic approaches described in the table are very similar to and compatible with those used in the QRIS Evaluation Toolkit (Lugo-Gil et al., 2011).

When reviewing the tables and the remainder of the Brief, it is helpful to be familiar with how three key QRIS terms – component, standard and indicator – are defined. The term quality **component** refers to the broad quality categories used in QRIS (such as staff qualifications, family engagement, and the learning environment). A quality **standard** is defined as a specific feature of quality such as specialized curriculum and assessment training in the staff qualifications component; a set of quality standards comprise each quality component**.** Quality **indicators** are metrics that can be measured or verified for each of the quality standards. A given quality standard could have one or multiple quality indicators that represent it in a QRIS. For example, in the category of staff qualifications, a standard may be "Teaching staff have specialized training in curriculum and assessment." An indicator related to this standard may be "At least 50% of teaching staff have completed the two-course statewide curriculum training session on curriculum and assessment."

Table 1. Four Related Approaches to Validating a QRIS

Approach	Activities and Purpose	Typical Questions Approach Addresses	Issues and Limitations
1. Examine the validity of key underlying concepts	Assess whether basic QRIS quality components and standards are the "right" ones by examining levels of empirical and expert support.	Do the quality components capture the key elements of quality? Is there sufficient empirical and expert support for including each standard?	Different QRISs may use different decision rules about what standards to include in the system.
2. Examine the measurement strategy and the psychometric properties of the measures used to assess quality	Examine whether the process used to document and verify each indicator is yielding accurate results. Examine properties of key quality measures, e.g., inter-rater reliability on observational measures, scoring of documentation, and inter-item correlations to determine if measures are psychometrically sound. Examine the relationships among the component measures to assess whether they are functioning as expected. Examine cut scores and combining rules to determine the most appropriate ways to combine measures of quality standards into summary ratings.	What is the reliability and accuracy of indicators assessed through program administrator self-report or by document review? What is the reliability and accuracy of indicators assessed through observation? Do quality measures perform as expected? (e.g., do subscales emerge as intended by the authors of the measures?) Do measures of similar standards relate more closely to each other than to other measures? Do measures relate to each other in ways consistent with theory? Do different cut scores produce better rating distributions (e.g., programs across all levels rather than programs at only one or two levels) or more meaningful distinctions among programs?	This validation activity is especially important given that some component measures were likely developed in low-stakes settings and have not been examined in the context of QRIS.[I]

Approach	Activities and Purpose	Typical Questions Approach Addresses	Issues and Limitations
3. Assess the outputs of the rating process	Examine variation and patterns of program-level ratings within and across program types to ensure that the ratings are functioning as intended. Examine relationship of program-level ratings to other quality indicators to determine if ratings are assessing quality in expected ways. Examine alternate cut points and rules to determine how well the ratings distinguish different levels of quality.	Do programs with different program-level ratings differ in meaningful ways on alternative quality measures? Do rating distributions vary by program type, e.g., ratings of center-based programs compared to ratings of home-based programs? Are current cut scores and combining rules producing appropriate distributions across rating levels?	These validation activities depend on a reasonable level of confidence about the quality components, standards and indicators as well as the process used to designate ratings.
4. Examine how ratings are associated with children's outcomes.	Examine the relationship between program-level ratings and selected child outcomes to determine whether higher program ratings are associated with better child outcomes.	Do children who attend higher-rated programs have greater gains in skills than children who attend lower-quality programs?	Appropriate demographic and program level control variables must be included in analyses to account for selection factors. Studies could be done on child and program samples to save resources. Findings do not permit attribution of causality about QRIS participation but inferences can be made about how quality influences children's outcomes.

Table 2. Data Needs, Data Sources and Analysis Methods for Selected Studies

Approach	Data needed	Data sources	Analysis methods
1. Examine the validity of key underlying concepts	Evidence about the relationship between key quality standards and desired outcomes. Expert opinions about proposed quality standards and indicators.	Empirical literature on how proposed components contribute to high quality care and improved child outcomes. Experts in early childhood education who can provide input on the quality standards and indicators.	Synthesis of available data relating to each component; Analysis of degree to which evidence meets criteria for relatedness; Consensus process; Decision rules that specify the value of components without an established evidence base.[II]
2. Examine the measurement strategies and psychometric properties of the measures used to assess quality.	Rating data from participating programs. Data from additional quality measures.	Most such data are collected as part of program ratings. Additional quality measures may be collected to allow comparisons with measures being used in the QRIS.	Distribution of provider scores on a given component; Correlations among components; Correlations of selected components with other measures.
3. Assess the outputs of the rating process	Program-level ratings from participating programs. Raw scores from measures of quality that are included in the rating. Data from additional quality measures that are not included in the rating.	Most of the necessary data are collected as part of program ratings. Another measure of quality may be administered to allow comparisons with program ratings.	Examination of rating distributions by program type; Correlations of program ratings with other measures; Changes in rating distributions using different cut scores.
4. Relate ratings to expected child outcomes.	Program rating data from participating programs. Assessments of child functioning.	Program rating data are collected as part of program ratings. Trained, reliable independent assessors collect data from individual children (may be a designated sample). Teacher reports on individual children.	Estimate the relationship between program ratings and child outcomes.[III]

Approach 1: Examine the validity of key underlying concepts. This approach involves examination of the elements or concepts that are to be included in program ratings. It is an important validation activity because it provides the foundation for the quality components, standards and indicators that together will produce program-level ratings and that will be the focus of quality improvement activities. Together, the components included in ratings, (e.g., staff qualifications, learning environment, family engagement) define quality for the QRIS. This validation activity provides justification and support for the elements of the QRIS. If the examination includes stakeholders, the process can also promote buy-in for the QRIS.

This validation approach asks whether quality components, standards and indicators included in a QRIS are the "right" ones, and is similar to what is proposed in the Toolkit, under *Validating Quality Standards* (Lugo-Gil et al., 2011). Because this effort addresses the cornerstone concepts and measures of the QRIS, it ideally would be conducted prior to the implementation of the QRIS.

For QRISs, the key concept is quality of care. The quality of care in early childhood education and school-aged care (ECE-SAC) programs is a complex, multi-dimensional construct; this complexity is amplified in centers by the fact that programs are comprised of multiple classrooms staffed by multiple individuals. Quality can be operationalized using a number of specific quality components. However, most QRISs have adopted similar ones. The QRIS Compendium found that six quality components were included in the majority of the 26 QRIS that were examined (Tout et al., 2010). These categories include licensing compliance (26 QRISs), classroom environment (24 QRISs), staff qualifications (26 QRISs), family partnership (24 QRISs), administration and management (23 QRISs) and accreditation (21 QRISs). Three categories—curriculum (14 QRISs), ratios and group size (13 QRISs), and child assessment (11 QRISs)—are included in half or just under half of the QRISs assessed. However, while similarities exist in the general quality components included in QRISs, the way in which each of these components of quality is measured varies substantially.

One activity that can help to validate a QRIS' underlying concepts involves assessing the degree to which the quality components in the QRIS rating include standards and indicators that have an empirical base linking them to key program, family and child outcomes. This assessment might include an examination of the degree to which each element as operationalized in the QRIS is viewed by experts as a valid measure of the component. A number of states (including Delaware, Rhode Island, Minnesota and Virginia) have used a systematic expert review process to help identify which quality components (and the standards and indicators that comprise each component) to include in their QRIS. Attention might also be paid to the views of programs and parents about the degree to which selected components reflect their priorities. For example, focus groups with parents were conducted in Minnesota to inform the development of the final rating tool used in the QRIS pilot (Minnesota Department of Education and Minnesota Department of Human Services, 2007)

Another activity which is part of this approach involves examining the research literature to determine the level of empirical support for each proposed component. This review would examine the research base on the proposed standards and indicators selected to represent program quality. The review would weigh the existing evidence and provide arguments for why a particular quality component should be included or excluded from the QRIS.

Purdue University's scientific review of the quality standards contained in Paths to Quality, Indiana's QRIS, demonstrates this approach. The overall goal of the review was to conduct an "external evaluation of the scientific validity" of the Paths to Quality standards (Elicker et al., 2007). The study included review of available evidence for the importance of each of the four quality components--Health and Safety, Learning Environment, Planned Curriculum, and National Accreditation-- and the relationship of the standards and indicators of each component to other measures of quality and to children's development and well-being. The review used standards of evidence to classify each proposed indicator. For example, one or two well-designed studies that supported the indicator was classified as "some evidence;" "substantial evidence" required more than five such studies. For three-quarters of the indicators, researchers found "substantial evidence" that they supported children's development.

Like many validation activities, such reviews ideally would be updated from time to time to determine if revisions to the QRIS would be advisable in light of new research findings. Such a review might utilize such tools as the *QRS Compendium* (Tout et al., 2010) or *Caring for Our Children* (AAP/APHA/NRC, 2011) as well as other recently published findings.

Approach 2: Examine the measurement strategies and the psychometric properties of the measures used to assess quality. A second type of validation effort focuses on the attributes of the individual measures in the QRIS as well as on the way in which the measures are combined to produce the summary rating of program quality. This approach is similar to what is discussed in the QRIS Evaluation Toolkit under *Validating the Construction of Quality Levels* (Lugo-Gil et al., 2011). This approach addresses how well the measures are working in the context of the QRIS. These efforts ask questions such as, "is there evidence that a given indicator measures what it purports to measure?" "If it claims to have a specific number of dimensions, do we find those dimensions in our data?" "Is there sufficient variance in scores on this indicator to justify its inclusion in the QRIS?" "Do scores on the indicator covary in expected ways with other measures of quality?"

Efforts to address these issues might involve an assessment of the distribution of participating provider scores on a given rating element. For example, in Zellman et al.'s (2008) evaluation of Colorado's QRIS, initial work revealed that the measure of family engagement then in use produced very little variation across programs; all programs achieved the highest score possible on this measure. This meant that the QRIS was expending substantial resources to collect data on a measure that did not differentiate among programs. Another validation activity might involve an assessment of the relationship of a given indicator to other indicators of quality, both those included in the QRIS and others. In such studies, it is important to look at the degree of correlation found: ideally, measures would be moderately correlated so that each measure provides some non-redundant program quality information (see Zellman et al., 2008 for an example). Correlation patterns also should make sense. For example, two measures of interaction quality should be more closely related to each other than to a measure of ratios. If such studies reveal for example that the correlation between ratios and interaction processes is very high, this result might argue for eliminating one or the other indicator from the QRIS, as they may not be providing additional information (although some QRISs include certain elements to ensure that they are paid attention to, even if their psychometric properties are not ideal).

The research literature provides limited guidance concerning the most appropriate ways to combine measures of quality elements into summary ratings (Lugo-Gil et al., 2011; Tout et al., 2009; Zellman et al., 2008). Yet this process is crucial to producing meaningful program quality ratings, which are the key output of the rating process. States that are collecting and combining data could use these data to conduct studies that examine the effects of altering cut scores or combination rules, much as Karoly and Zellman (2012) have done in a "virtual pilot" for California's QRIS, using data collected for another purpose, or as was done in studies in Minnesota (Tout et al., 2011) and Kentucky (Isner et al., 2012). These efforts will help QRIS designers and policy makers consider how well indicators are working, which indicators appear to be picking up variations in quality, and how closely different indicators relate to each other.

A number of other existing studies examine the properties of proposed QRIS indicators and can provide guidance to QRIS validation efforts (Scarr, Eisenberg, & Deater-Decker, 1994; Zellman & Perlman, 2008; Tout et al, 2011; McWayne & Melzi, 2011). Additionally, tools exist to help QRIS stakeholders review the options for QRIS measures and to support decision-making about the inclusion of new measures. For example, a Quality Measures Compendium is available and updated on a regular basis (Halle, Vick-Whittaker, & Anderson, 2010). If promising new measures are developed, it might be worthwhile to examine the performance of a new measure against the measure in current use.

Approach 3: Assess the outputs of the rating process. A third validation approach focuses on assessing the *outputs* of the rating system: the scores and levels that are assigned to providers who undergo a rating. Studies conducted under this approach examine the degree to which the quality levels in the QRIS are meaningfully distinct from each other. The results of these studies may indicate that measures, cut scores, or rules for combining measures need changing in order to distinguish quality levels effectively. Because these studies can result in proposals for significant changes to the composition of QRIS levels, it is helpful for these studies to occur prior to studies that examine associations between quality levels and children's development.

Output studies may focus on individual indicator scores, such as how providers score on an environmental rating, as well as on the program-level score that is the final output of the rating process. Studies conducted as part of this approach ask questions like, "are providers that received four stars actually providing higher quality care than those that earned three stars?" Studies using this approach may also address questions about cut scores, e.g., "do different cut scores produce dramatically different program-level ratings, and if so, which cut scores produce distributions that most closely relate to other measures of quality?" These studies typically rely on a measure of quality not included in the QRIS to make this assessment, and examine whether assessments on both measures vary in predictable ways.

The University of Southern Maine is conducting a validation study of Maine's QRIS to assess similarities and differences across program ratings; the study is also examining what if any differences exist between similar types of programs at different step levels (see Lahti et al., forthcoming, for further details on this study and several others.) For example, researchers in Maine administer the Environment Rating Scales (ERS; Harms & Clifford; 1989; Harms, Clifford & Cryer, 2005; Harms, Cryer & Clifford, 2006; Harms, Cryer & Clifford, 2007), which are not used to establish a rating in Maine's QRIS, and examine whether there are statistically significant differences in ERS scores between programs at different rating levels. These findings help program designers determine if the quality levels determined by QRIS ratings relate in expected ways to an external measure of global quality.

As a second example of validation studies using this approach, Karoly and Zellman (2012) used data collected for another purpose to model some of the features of a newly-designed California QRIS. The data come from a 2007 survey of center-based providers that is representative of the state. Observations were conducted in 251 centers serving children birth to 5. The purpose of this "virtual pilot" study was to determine the likely distribution of programs across QRIS tiers using specified cut points, examine the association among quality components, and to identify "outlier" quality elements on which otherwise well-rated programs tend to score poorly. This information is very valuable at the design phase; data on "outlier" elements is particularly helpful in understanding what it will take for programs to improve their rating in a QRIS that uses a block design to designate ratings (in which all indicators at one level must be met before a rating at the next level is possible). By examining such things as the relationship between scores on the Classroom Assessment Scoring System (CLASS; Pianta, La Paro & Hamre, 2008) and the Early Childhood Environment Rating Scale - Revised (ECERS-R; Harms, Clifford & Cryer, 2005), and the relationship between staff education and training and other measures of quality, the work can help policymakers assess the value of different measures of quality, provide input into establishing cut scores, and suggest targets for technical assistance efforts.

Other states also have conducted validation studies that focus closely on differences in QRIS levels. For example, Pennsylvania has studied programs participating in the Keystone STARS QRIS (Fiene, Greenberg, Bergsten, Fegley, Carl, & Gibbons, 2002; Barnard, Smith, Fiene, & Swanson, 2006; OCDEL (Office of Child Development and Early Learning), 2010; Manlove, Benson, Strickland, & Fiene, 2011) to determine if their program ratings were indicative of quality differentials across program types and services. Similarly, recent work in Indiana (Elicker, Langill, Ruprecht, Lewsader & Anderson, 2011) found that ERS scores varied with program-level ratings, while research in Minnesota found significantly higher scores on the ERS and CLASS only between the highest level (4-star) of the QRIS and the other rating levels (2- and 3-stars) (Tout et al., 2011). These findings are being used by program developers to make needed adjustments to quality indicators, metrics and cut scores.

Approach 4: Relate ratings to children's development. A fourth approach to validation focuses on children's development. It is similar to the Toolkit's *Linkages between quality levels and desired outcomes*, although it focuses more narrowly on child outcomes. For QRISs, the logic model asserts that higher quality care will be associated with better child outcomes. Therefore, one important piece of validation evidence concerns whether children make greater developmental gains in programs with higher program-level ratings than in programs with lower ratings.

Studies using this approach do not attempt to identify causal linkages between *QRIS participation* and children's outcomes. Instead, they examine whether the QRIS ratings and quality components that comprise the ratings are related in expected ways to measures of children's development. Appropriate designs and controls could allow causal inferences to be made about how *quality* (as measured and rated by the QRIS) influences children's outcomes.

To date, few QRIS validation studies have incorporated children's outcomes as they are costly and difficult to conduct. As Elicker and Thornburg (2011) note, results from such studies are mixed, at least in part because of the challenges of conducting them. A primary challenge is the inability to control for all the factors that may vary between children whose families have selected different programs. Additional challenges include recruitment of programs and children across all quality levels; availability of appropriate outcome measures for children of diverse ages, abilities, cultures and linguistic backgrounds; and, lack of variation in the quality of participating QRIS programs.

In Missouri, children who participated in programs with higher quality ratings showed significantly greater gains on measures of social-emotional development compared to children in programs with lower ratings (Thornburg et al., 2009). These effects were even more pronounced for low-income children. However, in an evaluation of Colorado's QRIS, linkages between the ratings and children's outcomes were not found (Zellman et al., 2008). Recent reports from Indiana (Elicker, Langill, Ruprecht, Lewsader, & Anderson, 2011) and Minnesota (Tout et al., 2011) found no consistent relationships between program ratings and measures of child outcomes. A number of possible explanations were offered for the lack of expected linkages, including overall low levels of quality in participating QRIS programs (perhaps not meeting a threshold of quality necessary to detect linkages with child outcomes; see Zaslow et al., 2010 for further discussion of quality thresholds) and a lack of variation among participating programs and families. Yet, even with these limitations, program administrators in both Indiana and Minnesota have used the findings to recommend changes to the structure and content of the QRIS.

Developing a Validation Plan

Given the complexity of validation, it is advisable to develop a plan for system validation as early as possible in the QRIS design process. Ideally, the validation plan will be part of a larger evaluation plan designed to address a wider range of important questions the answers to which will guide refinement of the QRIS and its implementation. The plan should include the key questions that will be addressed and the methods to be used to address each one. One advantage of developing a plan early is that it may highlight opportunities to conduct a number of the proposed efforts as part of the implementation of the QRIS itself or as part of planned evaluation activities. A comprehensive approach to validating a QRIS ideally will include studies under each of the four approaches described above. Table 3 outlines issues in the timing of validation studies, discusses their relative cost, and suggests strategies for addressing validation questions if resources do not permit the implementation of validation studies.

Table 3. Considerations in Developing a Validation Plan

Approach	Timing and Duration	Cost considerations	Options to consider[IV]
1. Examine the validity of key underlying concepts	Ideally conducted prior to QRIS implementation. Study should be able to be completed within 3-6 months.	Relatively inexpensive. This work can be contracted to a local university, consultant or research firm.	Many states are using similar concepts and measures; their efforts will provide useful information.[V]
2. Examine the measurement strategies and psychometric properties of the measures used to assess quality	Must wait until ratings are implemented, although individual measures themselves might be available from other sources and could be examined earlier.[VI]	Depends on data quality and amount of analysis. Additional measures will increase costs, particularly if the measure is observational.	Can rely to some extent on existing research on each of the components. Consider using available data for a "virtual pilot."[VII]
3. Assess the outputs of the rating process	Must wait until ratings are implemented. Once data are available, several studies could be conducted using the same data set.	Depends on data quality and amount of analysis. Additional measures will increase costs, particularly if the measure is observational.	This work is state system-dependent so is not readily borrowed, though lessons learned about structure and cut-points can be shared across QRISs.
4. Relate ratings to children's development	Best to launch these studies when the QRIS rating process is stable and adequate numbers of programs have been rated.	Costs for the collection of child data are very high. Study could be done just with one cohort of children and two rounds of data collection (fall and spring) to assess developmental gains.	Requires significant funds, a powerful research design, and research expertise. Sampling children and programs will substantially reduce costs.

Summary and Conclusions

Validation is a complex, ongoing, iterative process. The objective of validation activities is to understand whether the rating process is able to distinguish among programs of different quality levels and whether program ratings are associated in meaningful ways to children's outcomes.

Validation activities help to determine whether key design decisions are working well in practice. States and localities that have implemented QRISs are expending substantial resources to train raters, fund ratings, support various forms of technical assistance, and provide a range of improvement incentives. All of these efforts assume that the ratings are accurate and the system is performing as intended. QRIS design decisions often rely heavily on the judgments of experts and on colleagues in other states, because there is limited empirical data on which to base them. For this reason, it is critical for states to set in place a process for assessing how well the design decisions underlying the system are working. Validation activities do this.

Ideally, validation is an ongoing process based on a carefully designed validation plan. The plan should include all four validation approaches, although resource constraints may limit these efforts, and may particularly limit studies that include child outcomes. A good validation plan, thoughtfully developed and implemented, can provide information critical to improving the system at many points in the process, and increase the odds of its ultimate success. Validation is unquestionably challenging, but no more so than the launch and operation of a QRIS or its evaluation. The networks and references in the next section can help states develop a deeper understanding of validation approaches and help them construct and implement validation plans that address stakeholder and system needs and produce timely and valuable information.

Resources and References

Resources

INQUIRE – Quality Initiatives Research and Evaluation Consortium

http://www.acf.hhs.gov/programs/opre/cc/childcare_technical/index.html

The purpose of INQUIRE is to support high quality, policy-relevant research and evaluation on Quality Rating and Improvement Systems and other quality initiatives by providing a learning community and resources to support researchers and evaluators. INQUIRE also provides input and information to state administrators and other policymakers and practitioners on evaluation strategies, new research, interpretation of research results, and implications of research for practice. Research briefs are available on topics related to QRIS evaluation issues and strategies.

CCEERC – Child Care and Early Education Resource Connections

http://www.childcareresearch.org/ search under Quality Rating and Improvement Systems.

This site has many additional reports and resources, such as:

Quality Rating Systems: A Key Topic Resource List. New York: Child Care & Early Education Research Connections.

http://www.researchconnections.org/files/childcare/keytopcis/QualityRatingSystems.pdf

This resource list is an annotated bibliography of selected research focused on the design, implementation, and evaluation of Quality Rating Systems and Quality Rating and Improvement Systems in early childhood and after school settings.

The Child Care Quality Rating System (QRS) Assessment

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G. & Boller, K. (2010). ***The Child Care Quality Rating System (QRS) Assessment: Compendium of Quality Rating Systems and Evaluations, OPRE Report.*** Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

http://www.acf.hhs.gov/programs/opre/cc/childcare_quality/compendium_qrs/qrs_compendium_final.pdf

Describing 26 Quality Rating Systems nationwide (19 statewide and 7 local or pilot), the Compendium presents comprehensive information through cross-QRS matrices and individual QRS profiles.

Lugo-Gil, J., Sattar, S., Boss, C., Boller, K. Tout, K., & Kirby, G. (2011). ***The Quality Rating and Improvement System (QRIS) Evaluation Toolkit. OPRE Report #2011-31.*** Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

http://www.acf.hhs.gov/programs/opre/cc/childcare_quality/qris_toolkit/qris_toolkit.pdf

The QRS Assessment Toolkit will provide guidance, recommendations and evaluation support on a range of topics including: development of a logic model and research questions, evaluation design and methods, and selection of measures.

QRIS National Learning Network

http://qrisnetwork.org/

The Network provides information, learning opportunities, and direct technical assistance to states that have a QRIS or that are interested in developing one. Its National Resource Library assists states in learning more about QRIS and their elements and in QRIS planning. The library contains, toolkits, handouts and published documents on a variety of searchable topic areas.

The Networks' State Resource Library contains detailed QRIS implementation information, including training guides, forms, and technical assistance materials that individual states have developed for their QRIS.

State QRIS Contacts who have agreed to serve as peer resources for one another are listed, as are Technical Assistance Providers.

Additional Resources

Lahti, M., Langill, C., Sabol, T., Starr, R., & Tout, K., (in progress). ***Validating Standards in Child Care Quality Rating and Improvement Systems: Exploring Validation Activities in Four States, OPRE Report.*** Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This report will provide case studies of four states that have undertaken validation studies in their respective states. This report provides validation and evaluation approaches, identification of similar QRIS standards amongst the four states, description of cross case analysis QRIS validity issues and the results of the validation conceptual model from this brief examining the following: concepts of quality, measures used to assess quality, outputs or scores of the rating process, and if ratings are related to expected outcomes. It is the companion document to supplement this guide in which four states validation experiences are highlighted.

Halle, T., Vick Whittaker, J. E., & Anderson, R. (2010). ***Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition***. Washington, DC: Child Trends. Prepared by Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
http://www.acf.hhs.gov/programs/opre/cc/childcare_technical/reports/complete_compendium_full.pdf

The Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition was compiled by Child Trends for the Office of Planning, Research and Evaluation of the Administration for Children and Families, U.S. Department of Health and Human Services, to provide a consistent framework with which to review the existing measures of the quality of early care and education settings. The aim is to provide uniform information about quality measures. It is hoped that such information will be useful to researchers and practitioners, and help to inform the measurement of quality for policy-related purposes.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Academy of Pediatrics, American Public Health Association, & National Resource Center for Health and Safety in Child Care (AAP/APHA/NRCHSCC) (2011). *Caring for our children: National health and safety performance standards guidelines for early care and education programs*. Elk Grove Village, Illinois: American Academy of Pediatrics.

Anastasi, A. (1986). *Psychological testing (5th ed.).* NY: Macmillan.

Barnard, W., Smith, W., Fiene, R., & Swanson, K. (2006). *Evaluation of Pennsylvania's Keystone STARS quality rating system in child care settings*. Pittsburgh, Pennsylvania: University of Pittsburgh Office of Child Development.

Cizek, Gregory J. Introduction to Validity. Presentation to the National Assessment Governing Board of NAEP. August, 2007.

Elicker, J., Langill, C., Ruprecht, K., & Kwon, K. (2007). *Paths to quality: A child care quality rating system for Indiana: What is its scientific base*. West Lafayette, IN: Purdue University.

Elicker, J., & Thornburg, K. (2011). *Evaluation of quality rating and improvement systems in early childhood programs and school age care: measuring children's development, research to policy, research to practice brief OPRE 2011-11c*. Washington, DC: Department of Health and Human Services, Administration of Children and Families, Office of Planning, Research, and Evaluation.

Elicker, J., Langill, C.C., Ruprecht, K., Lewsader, J., & Anderson, T. (2011). Evaluation of "Paths to QUALITY", Indiana's child care quality rating and improvement system. West Lafayette, IN: Purdue University.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179-197.

Fiene, R., Greenberg, M., Bergsten, M., Fegley, C., Carl, B., & Gibbons, E. (2002). *The Pennsylvania early childhood quality settings study*. Harrisburg, Pennsylvania: Governor's Task Force on Early Childhood.

Guion, R.M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement*, 1, 1-10.

Halle, T., Vick-Whittaker, J. E., & Anderson, R. (2010). *Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition*. Washington, DC: Child Trends. Prepared by Child Trends for the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Harms, T., & Clifford, R. (1989). *Family Day Care Environmental Rating Scale*. New York: Columbia University Teachers College Press.

Harms, T., Clifford, R., & Cryer, D. (2005). *Early Childhood Environmental Rating Scale – Revised*. New York: Columbia University Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2006). *Infant Toddler Environmental Rating Scale – Revised*. New York: Columbia University Teachers College Press.

Harms, T., Cryer, D. & Clifford, R. (2007*). Family Day Care Environmental Rating Scale – Revised*. New York: Columbia University Teachers College Press.

Isner, T., Soli, M., Rothenberg, L., Moodie, S., & Tout, K. (2012). *Alternative rating structures for Kentucky STARS for KIDS NOW, Evaluation Brief #6*. Washington, D.C.: Child Trends.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Education Measurement*, 38, 319-42.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.) *Educational Measurement, 4th edition* (pp. 17-64). Westport, CT: Praeger.

Karoly, L. A. and Zellman, G.L. (2012). *How Would Programs Rate Under California's Proposed Quality Rating and Improvement System? Evidence from Statewide and County Data on Early Care and Education Program Quality*. Santa Monica, CA: RAND Corporation.

Keyes, C. (2002). A way of thinking about parent/teacher partnerships for teachers. *International Journal of Early Years Education, 10*(3), 177 – 191.

Kontos, S. (1987). The attitudinal context of family day care relationships. In D. Peters & S. Kontos (Eds.), *Continuity and discontinuity of experience in child care* (pp. 91 – 113). Norwood, NJ: Ablex Publishing.

Lugo-Gil, J., Sattar, S., Boss, C., Boller, K., Tout, K., & Kirby, G. (2011). *The Quality Rating and Improvement System (QRIS) Evaluation Toolkit. OPRE Report #2011-31.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation.

Manlove, E., Benson, M., Strickland, M., & Fiene, R. (2011). *A comparison of regulated child care in rural and urban Pennsylvania*. Harrisburg, Pennsylvania: The Center for Rural Pennsylvania.

McGrath, W. (2007). Ambivalent partners: Power, trust, and partnerships in relationships between mothers and teachers in a full-time child care center. *Teachers College Record,* 109(6), 1401 – 1422.

McWayne, C. & Melzi, G., (2011). Family engagement during preschool, paper presented to the Head Start Advisory Committee on Research and Evaluation, Washington DC.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychology*, 30, 955-966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012-1027.

Minnesota Department of Education and Minnesota Department of Human Services. (January, 2007). *Child care information and rating system – parent focus group results.* DHS-4965-ENG 1-07. St. Paul, MN.

Office of Child Development and Early Learning (OCDEL) (2010). *Keystone STARS Program Report*. Harrisburg, Pennsylvania: Department of Public Welfare.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment and Scoring System (CLASS).* Baltimore, MD: Paul H. Brookes Publishing Co, Inc.

Scarr, S., Eisenberg, M., & Deater-Deckard, K. (1994). Measurement of quality in child care centers. *Early Childhood Research Quarterly*, 9, 131-151.

Shimoni, R. (1992) Parent involvement in early childhood education and day care, *Sociological Studies of Child Development,* 5, 73–95.

Tenopyr, M.L. (1977). Content-construct confusion. *Personnel Psychology*, 30, 47-54.

Thornburg, K., Mayfield, W.A., Hawks, J.S., & Fuger, K.L. (2009). *The Missouri Quality Rating System School Readiness Study*. University of Missouri--Columbia. Center for Family Policy and Research.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). *Issues for the Next Decade of Quality Rating and Improvement Systems,* OPRE Issue Brief. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G. & Boller, K. (2010). *Compendium of Quality Rating Systems and Evaluations.* OPRE Report. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Tout, K., Starr, R., Isner, T., Cleveland, J., Albertson-Junkans, L., Soli, M., & Quinn, K. (2011). *Evaluation of Parent Aware: Minnesota's Quality Rating and Improvement System Pilot, Final Evaluation Report*. Produced for the Minnesota Early Learning Foundation. Minneapolis, MN: Child Trends.

Zaslow, M., Anderson, R., Redd, Z., Wessel, J., Tarullo, L. & Burchinal, M. (2010). *Quality Dosage, Thresholds, and Features in Early Childhood Settings: A Review of the Literature*, OPRE 2011-5. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zellman, G.L., Brandon, R.N., Boller, K., & Kreader, J.L. (2011). *Effective evaluation of quality rating and improvement systems for early care and education and school-age care, Research-to-Policy, Research-to-Practice Brief* OPRE 2011-11a. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Zellman, G. L., & Karoly, L.A. (2012). *Incorporating Child Assessments into State Early Childhood Quality Improvement Initiatives.* Santa Monica, CA: RAND Corporation.

Zellman, G.L., & Karoly, L.A. (2012). *Moving to Outcomes: Approaches to Incorporating Child Assessments into State Early Childhood Quality Rating and Improvement Systems.* Santa Monica, CA: RAND Corporation, OP-364-PF.

Zellman, G.L. & Perlman, M. (2006). Parent involvement in child care settings: Conceptual and measurement issues. *Early Child Development and Care*, 176(5), 521-538.

Zellman, G. L., & Perlman, M. (2008). *Child-Care Quality Rating and Improvement Systems in Five Pioneer States: Implementation Issues and Lessons Learned*. Santa Monica, CA: RAND Corporation.

Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar Early Learning quality rating and improvement system as a tool for improving child-care quality*. (MG-650-QEL). Santa Monica, CA: RAND Corporation.

Endnotes

[I] Validity is not attached to a measure, but to a measure used for a particular purpose in a particular context. This means that measures which may be valid for one use must be validated again for use in a different context (AERA, APA, & NCME, 1999). Measures developed in low-stakes contexts, e.g., for use in research or program self-assessments, must be validated again in high-stakes contexts because those being assessed may react in high-stakes contexts in ways that could undermine the meaningfulness of interpretations derived from those measures (AERA, APA, & NCME, 1999).

[II] Some components such as parent involvement have been included in QRISs even when strong empirical support of the ability of measures to distinguish among programs of different quality was lacking because designers believed that if they were not, programs would ignore these components in favor of measured ones.

[III] Random assignment of children to programs with different quality ratings is not possible in QRIS. Alternative analytic approaches must be used that employ adequate controls for selection bias. See Zellman and Karoly (2012) for further discussion of this approach.

[IV] This column recognizes that state budgets are limited and validation is rarely seen as the highest priority. Ideally, states might combine data and efforts to conduct some of these studies.

[V] Ideally, states might combine data and efforts to conduct some of these studies.

[VI] However, as noted above, measures collected in low-stakes and high-stakes settings cannot be assumed to be comparable.

[VII] It may be possible to use existing data to test assumptions and measures. See, for example, Karoly and Zellman (2012), for a description of such work in California.